Effect Analysis of Factors in Generalized Linear Models

Nobuoki Eshima

Department of Medical Information Analysis, Faculty of Medicine,
Oita Medical University, Oita 879-5593, Japan
e-mail: eshima@oita-med.ac.jp

Abstract

Effect analysis of factors on responses or measurements in generalized linear models is important in many application studies, e.g. medicine, sociology, psychology etc. The effect analysis is usually made in testing the main effects and interactions for levels of factors. It is significant to consider summary measures of the effects of factors, i.e. contributions of factors on explaining a response variable. In the present paper, we discuss the partial, association, and total effects of factors in generalized linear models with canonical links. The summary effect measures are provided through a discussion of log odds ratio, and it is shown that the effect measures are related to the Kullback-Leibler information. The present approach is applied to some important generalized linear models.

Key Words: Association effect; Effect analysis; Generalized linear model; Log odds ratio; Partial effect; Total effect.

1. Introduction

Analyses of the effects of factors on responses or measurements have been made in many scientific fields, e.g. medicine, sociology, psychology etc. Generalized linear models (Nelder and Wedderburn, 1972; McCullagh and Nelder, 1989) are popularly used for describing regression relationships between response variables and factors in both experimental and observational studies. The generalized linear models (GLMs) are an exponential family of distributions that include various important regression models, e.g. ordinary linear regression model (Draper and Smith, 1966), logistic regression (logit) model (Cox and Snell, 1989; Hosmer and Lemeshow, 1989), Poisson regression model (Vonesh, 1990; Christiansen and Moris, 1997), Loglinear model (Haberman, 1979) etc. Generalized linear models are also applied to statistical quality control (Hamada and Nelder, 1997), because there are many cases of nonnormal responses in industrial experiments (Lewis, et al. 2001a). The usual least square approach with response transformation and the GLM approach are compared, and Lewis et al. (2001a, b) investigated advantages of the latter approach.

The usual effect analysis with GLMs is to statistically test the main effects and the interactions of levels of factors, and to interpret the results. After the analysis, it is important to summarize the effects of the factors, that is, comparison of contributions of factors on the variation of response Y. In this paper, the partial, association and total effects of factors are discussed for generalized linear models. The effect measures for the partial, association and total effects are derived from a discussion of log odds ratio, and properties of the effect measures are considered. Examples of

GLMs are also provided.

2. Basic Idea of Measuring the Effects of Factors in GLMs with Canonical Links

Let $\boldsymbol{X} = (X_1,X_2,\ldots,X_K)^{\mathrm{T}}$ be a $K\times 1$ factor vector; let Y be a response variable; and let $f(y|\boldsymbol{x})$ be the conditional probability or density function of Y given $\boldsymbol{X} = \boldsymbol{x} = (x_1,x_2,\ldots,x_K)^{\mathrm{T}}$. Then, a GLM with the canonical link is described as follows:

$$f(y|\boldsymbol{x}) = \exp\{(y\,\theta - b(\theta))/a(\phi) + c(y,\phi)\}, \tag{2.1}$$

where θ and ϕ are parameters and for $\boldsymbol{\beta} = (\beta_1, \beta_2,\ldots, \beta_K)^{\mathrm{T}}$

$$\theta = \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x}. \tag{2.2}$$

Without loss of generality, we set $a(\phi) > 0$. The effect of factor $X_i = x_i$ may be defined by β_i from an analogy to the ordinary linear regression model, because the predictor (2.2) increases by β_i for unit change of the factor. However, it is a question how this effect is interpreted except the usual regression model. Let $\mathrm{OR}(y,y^*|x,x^*)$ be the odds ratio with respect to x, x^*, y, and y^*. Then, we have

$$\log \mathrm{OR}(y,y^*|\boldsymbol{x},\boldsymbol{x}^*) = (y - y^*)\,\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*)/a(\phi). \tag{2.3}$$

The first interpretation of this quantity is that (2.3) is the inner product of the response $y - y^*$ and the predictor $\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x} - \boldsymbol{x}^*)$ with respect to $a(\phi)$. The second interpretation is as follows. The conditional log odds of $Y = y$ over $Y = y^*$ given $\boldsymbol{X} = \boldsymbol{x}$ is

$$\log\{f(y|\boldsymbol{x})/f(y^*|\boldsymbol{x})\} = \{-\log f(y^*|\boldsymbol{x})\} - \{-\log f(y|\boldsymbol{x})\}$$

= the amount of the conditional uncertainty of $Y = y^*$ minus that of $Y = y$ given $\boldsymbol{X} = \boldsymbol{x}$.

From this, the above quantity is a decrease of uncertainty of y over y^*. Hence, the log odds ratio (2.3) can be interpreted as the change of uncertainty of y over y^* for the change of factors from $\boldsymbol{x}^*$ to $\boldsymbol{x}$. When we substitute the baselines x^* and y^* in (2.3) for the expectations μ_X and μ_Y, respectively, we get

$$\log \mathrm{OR}(y, \mu_Y|\boldsymbol{x}, \mu_X) = (y - \mu_Y)\,\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{x} - \mu_X)/a(\phi). \tag{2.4}$$

This odds ratio can be viewed as the total effect of $\boldsymbol{X} = \boldsymbol{x}$ on $Y = y$. In order to summarize the above quantity, we define the total effect of $\boldsymbol{X}$ on Y by

$$e_{\mathrm{T}}(\boldsymbol{X}\to\boldsymbol{Y}) = \mathrm{E}\{(Y - \mu_Y)\,\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{X} - \mu_X)/\,a(\phi)\}. \tag{2.5}$$

This quantity is the covariance of the response Y and the predictor $\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{X}$ with respect to $a(\phi)$. For example, the following linear regression model is considered:

$$Y = \mu + \boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{x} + e, \tag{2.6}$$

where error term e is distributed according to normal distribution $N(0, \sigma^2)$. In this case, the dispersion parameter is $a(\phi) = \sigma^2$. From this model, we have

$$e_{\mathrm{T}}(\boldsymbol{X}\to\boldsymbol{Y}) = \mathrm{E}\{(\boldsymbol{X} - \mu_X)^{\mathrm{T}}\,\boldsymbol{\beta}\,\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{X} - \mu_X)\}/\sigma^2 = (\mathrm{Var}(\mathrm{Y}) - \sigma^2)/\sigma^2.$$

This is the ratio of the explained variation of Y by factor vector $\boldsymbol{X}$ to the variation of the error term. Since $a(\phi)$ in (2.5) is a dispersion parameter of the GLM (2.1), the parameter is referred to as a generalized dispersion of the GLM (2.1), i.e. the generalized dispersion of the error of the prediction. The numerator in (2.5)

$$\mathrm{E}\{(Y - \mu_Y)\,\boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{X} - \mu_X)\}$$

can be viewed as a generalized explained dispersion of response Y by the predictor vector $\boldsymbol{X}$. Hence,

the total effect (2.5) can be interpreted as the ratio of a generalized explained variation of response Y by factor $\boldsymbol{X}$ to a generalized error variation of the GLM (2.1).

Second, we consider the partial effects of X_i on Y. Let $\boldsymbol{X}^{(i)} = (X_1,X_2,..,X_{i-1},X_{i+1},\ldots,X_K)^{\mathrm{T}}$; let μ

$_i(\boldsymbol{x}^{(i)})$ be the conditional expectation of X_i given $\boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)}$; let $\boldsymbol{\mu}_{\boldsymbol{X}}(\boldsymbol{x}^{(i)}) = (x_1, x_2, .., x_{i-1}, \mu_i(x^{(i)}), x_{i+1}, \ldots, x_K)^{\mathrm{T}}$, i.e. the conditional expectation of $\boldsymbol{X}$ given $\boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)}$; and let $\mu_{\boldsymbol{Y}}(\boldsymbol{x}^{(i)})$ be the conditional expectation of Y given $\boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)}$. Then, we get

$$\log \mathrm{OR}(y, \mu_{\boldsymbol{Y}}(\boldsymbol{x}^{(i)}) | \boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}(\boldsymbol{x}^{(i)})) = (y - \mu_{\boldsymbol{Y}}(\boldsymbol{x}^{(i)}))\, \beta_i (x_i - \mu_i(\boldsymbol{x}^{(i)}))/a(\phi). \tag{2.7}$$

The quantity (2.7) can be viewed as the direct effect of level $X_i = x_i$ on $Y = y$. By taking the conditional expectation with respect to X_i and Y given $\boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)}$, we have

$$\mathrm{E}\{\log \mathrm{OR}(y, \mu_{\boldsymbol{Y}}(\boldsymbol{x}^{(i)}) | \boldsymbol{x}, \boldsymbol{\mu}_{\boldsymbol{X}}(\boldsymbol{x}^{(i)}) | \boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)})\} = \beta_i \mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)})/a(\phi), \tag{2.8}$$

where $\mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)})$ are the conditional covariance of Y and X_i given $\boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)}$. The partial effect of X_i on Y is defined by the expectation of the above quantity with respect to $\boldsymbol{X}^{(i)}$:

$$\mathrm{e}_{\mathrm{P}}(X_i \rightarrow Y) = \beta_i \mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)})/a(\phi), \tag{2.9}$$

where $\mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)})$ is the expectation of $\mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)})$ with respect to $\boldsymbol{X}^{(i)}$, i.e. for the marginal density or probability function of $\boldsymbol{X}^{(i)}$, $g(\boldsymbol{x}^{(i)})$,

$$\mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)}) = \int \mathrm{Cov}(Y, X_i | \boldsymbol{X}^{(i)} = \boldsymbol{x}^{(i)}) g(\boldsymbol{x}^{(i)}) \mathrm{d}\boldsymbol{x}^{(i)}.$$

The partial effect (2.7) can be interpreted as the ratio of the generalized partially explained variation of response Y by factor X_i to a generalized error variation of the GLM (2.1). For example, in an ordinary linear regression model (2.6) we have

$$\mathrm{e}_{\mathrm{P}}(X_i \rightarrow Y) = \beta_i^2 \mathrm{Var}(X_i | \boldsymbol{X}^{(i)})/\sigma^2.$$

Third, the total effects of factors are discussed. According to the above discussion, the direct effect of factor vector $\boldsymbol{X}^{(i)}$ is defined by

$$\mathrm{e}_{\mathrm{P}}(\boldsymbol{X}^{(i)} \rightarrow Y) = \sum_{j \neq i} \beta_j \mathrm{Cov}(Y, X_j | X_i)/a(\phi). \tag{2.10}$$

The total effect of factor X_j is defined by

$$\begin{aligned} \mathrm{e}_{\mathrm{T}}(X_i \rightarrow Y) &= \mathrm{e}_{\mathrm{T}}(\boldsymbol{X} \rightarrow \boldsymbol{Y}) - \mathrm{e}_{\mathrm{P}}(\boldsymbol{X}^{(i)} \rightarrow Y) \\ &= \beta_i \mathrm{Cov}(Y, X_i)/a(\phi) + \sum_{j \neq i} \beta_j \mathrm{Cov}(\mu_j(X_i), \mu_Y(X_i))/a(\phi). \end{aligned} \tag{2.11}$$

The total effect is interpreted as the inner product of $y - \mu_{\boldsymbol{Y}}$ and the predictor

$$\beta_i (x_i - \mu_i)/a(\phi) + \sum_{j \neq i} \beta_j (\mu_j(x_i) - \mu_j)/a(\phi).$$

By subtracting the partial effect of X_i (2.9) from the total effect (2.11), we define the association effect of X_i by

$$\begin{aligned} \mathrm{e}_{\mathrm{A}}(X_i \rightarrow Y) &= \mathrm{e}_{\mathrm{T}}(X_i \rightarrow Y) - \mathrm{e}_{\mathrm{P}}(X_i \rightarrow Y) \\ &= \beta_i \mathrm{Cov}(\mu_i(\boldsymbol{X}^{(i)}), \mu_{\boldsymbol{Y}}(\boldsymbol{X}^{(i)}))/a(\phi) + \sum_{j \neq i} \beta_j \mathrm{Cov}(\mu_j(X_i), \mu_Y(X_i))/a(\phi). \end{aligned} \tag{2.12}$$

The above effect is interpreted as the effect that is made by the association between X_i and $\boldsymbol{X}^{(i)}$. If X_i is independent of $\boldsymbol{X}^{(i)}$,

$$\mathrm{e}_{\mathrm{A}}(X_i \rightarrow Y) = 0.$$

The total effect is

$$\mathrm{e}_{\mathrm{T}}(X_i \rightarrow Y) = \mathrm{e}_{\mathrm{P}}(X_i \rightarrow Y) + \mathrm{e}_{\mathrm{A}}(X_i \rightarrow Y).$$

By replacing the factor X_i in the above discussion by the set of factors, the total, direct and indirect effects of the set of factors can be defined. The following quantity is the contribution proportion of the effect of factor X_i on Y:

$$\mathrm{C}(X_i) = \mathrm{e}_{\mathrm{T}}(X_i \rightarrow Y)/\mathrm{e}_{\mathrm{T}}(\boldsymbol{X} \rightarrow \boldsymbol{Y}). \tag{2.13}$$

It is seen that the effects defined above are scale-invariant. If X_i (i =1,2,…,K) are mutually independent, from (2.5) we have

$$e_T(\boldsymbol{X} \to \boldsymbol{Y}) = \sum_{i=1}^{K} e_P(X_i \to Y) = \sum_{i=1}^{K} e_T(X_i \to Y). \quad (2.14)$$

From the above consideration, we have

$$\begin{aligned} e_T(\boldsymbol{X} \to Y) &= e_T(X_i \to Y) + e_P(\boldsymbol{X}^{(i)} \to Y) \\ &= e_P(X_i \to Y) + e_T(\boldsymbol{X}^{(i)} \to Y). \end{aligned}$$

3. Application to GLMs

In this section, the above idea is applied to normal distributions and the logit model. The details are omitted.

4. Numerical Examples

The data concerning the effects of AZT in showing the development of AIDS sysamptoms (Agresti, 1996, p.119) and the data in a study of length of time spent on individual home visit by public nurses (Daniel, 1999, pp. 348-353) are reanalyzed with the present approach. The details are omitted.

REFERENCES

Agresti, A. (1990). *Categorical Data Analysis*, New York: John Wiley & Sons, Inc.

Agresti, A. (1996). *An Introduction to Categorical Data Analysis*, New York: Johm Wiley & Sons, Inc.

Christiansen, C. L. and Morris, C. N. (1997). Hierarchical Poisson regression modeling, *Journal of the American Statistical Association*, **92**, 618-632.

Cox, D. R. and Snell, E. J. (1989). *Analysis of Binary Data*, London: Chapman and Hall.

Daniel, W. W. (1999). *Biostatistics: A Foundation for Analysis in the Health Sciences*, Seventh Edition, New York; John Wiley & Sons, Inc.

Drayper, N. R. and Smith, H. (1966). *Applied Regression Analysis*, New York: John Wiley & Sons, Inc.

Haberman, S. J. (1979). *Analysis of Qualitative Data, Vol. 2*, New York: Academic Press, Inc.

Hamada, M. and Nelder, J. A. (1997). Generalized linear models for quality-improvement experiments, *Journal of Quality Technology*, 29, 292-304.

Hosmer, D. and Lemeshow, S. (1990). *Applied Logistic Regression*, New York; John Wiley & Sons, Inc.

Lewis, S. L., Montgomery, D. C., and Myers, R. H. (2001). Confidence interval coverage for designed experiments analyzed with generalized linear models, *Journal of Quality Technology* (to appear).

Lewis, S. L., Montgomery, D. C., and Myers, R. H. (2001). Examples of designed experiments with nonnormal responses, *Journal of Quality Technology* (to appear).

McCullagh, P. and Nelder, J.A. (1989) *Generalized Linear models*, 2nd ed. London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R.W.M. (1972) Generalized linear model, *Journal of the Royal Statistical Society*, **A**, 135, 370-384.

Vonesh, E. F. (1990). Modelling Peritonitis rates and associated risk factors for individuals on continuous ambulatory dialysis, *Statistics in Medicine*, 9, 263-271.